Appendix 2. Star Schema Example

A star schema (**figure 2-1**) is a DSS construction that arranges data in a format that facilitates analysis. Within a data mart, several star schemas may be constructed to answer various business questions. There are two components that must be defined when creating a star schema: facts and dimensions. For each component, at least two key issues must be addressed: goals and influences. The text that follows drills down to actual issues.

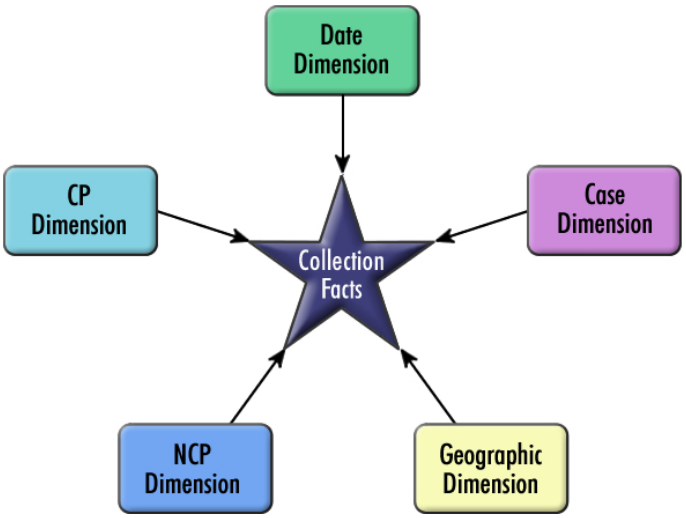

Figure 2-1. Collections Star Schema

Facts

For example, if we want to create a star schema for the purpose of organizing data to analyze collections (our goal), we will need to define what data we need (payments, both collections and arrears) and identify the sources of this data (influences).

Dimensions

Dimensions are the elements you want to measure the facts against. In our collections example, dimensions are time, geography, demography, and occupation. These dimensions address questions like:

- What were collections last year, by quarter?
- What were collections in the third quarter of this year, by occupation?
- What was the average time from case establishment to the first collection last year?

Dimension Tables

Dimension data is arranged in tables. **Figure 2-2** illustrates the step-by-step, logical method for developing dimensional data for incorporation into a data mart. Ignoring any step invites disaster.

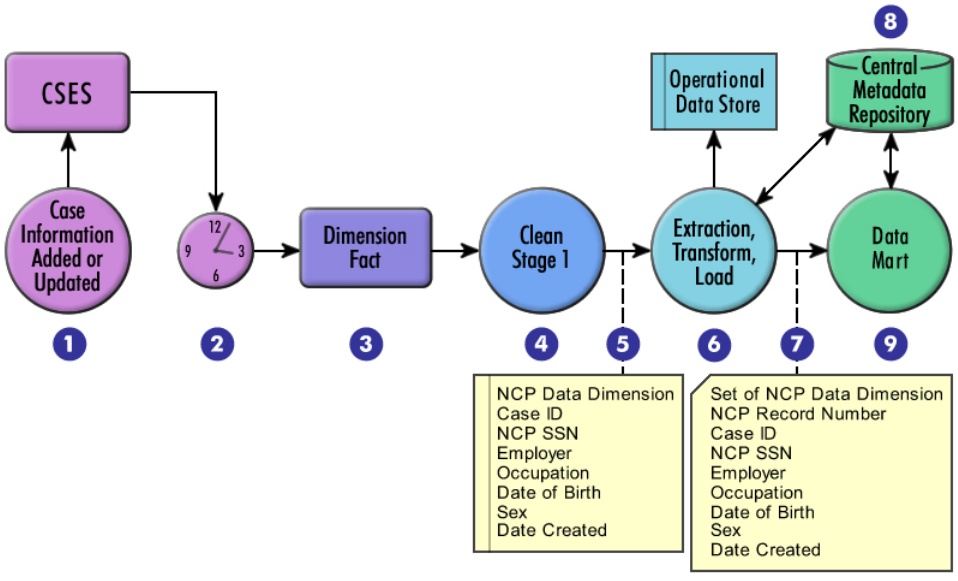

Figure 2-2. Developing Dimensional Data for Incorporation

Step 1 represents the day-to-day operations of CSES. New clients are added, current clients are updated, and clients are removed. All of this information needs to be included in the data mart (steps 3 – 9). If a client is removed from CSES, it is important to ensure transfer of the information about the closed cases to the data mart, especially when you are first building an SCS-DSS.

Step 2 represents a time delay between when information is input into CSES and copied into the SCS-DSS. It is neither practical nor efficient to immediately transfer CSE dimension information:

- Using near real-time data does not significantly impair SCS-DSS results.
- By allowing for a 24-hour delay, you can minimize the amount of bad data and reduce the workload on CSES. Often CSES inputs are corrected within 24 hours of initial entry.

The actual transfer schedule should be determined as part of the overall system design.

Step 3 is a representation of all the dimensional data that will be copied into the SCS-DSS. There are two approaches to developing this data:

- Take a complete snapshot of each case record and let the SCS-DSS's ETL process sort it out.
- Tailor the information required for the dimension table being populated.

There are advantages and disadvantages to both methods. These are shown in the table below.

	Snapshot of Each Case	Tailored Information
Advantages	All information is available. Easier to program. More adaptable to changes within the SCS-DSS.	Maintains focus on relevant business questions. Less processing effort required during ETL.
Disadvantages	Higher processing load on the SCS-DSS. Easier to lose data mart's focus.	Higher system load on CSES Data elements will be forgotten or new elements will be added, necessitating additional CSES programming.

Step 4 is the initial data-cleansing phase, during which dimension records are examined for compliance with existing CSES processing rules. The system rejects any records that do not conform to these rules and creates a log entry. Records that would be rejected are those with improper data types (text in a numeric field, for example), improper data (a 6-digit phone number or an 8-digit SSN), or malformed records (too long or too short). Edit checks within CSES should keep rejection rates low.

Step 5 represents a single cleansed NCP dimension record. All fields conform to CSES business rules.

Step 6 is the ETL process, which standardizes data to meet the data mart's requirements. Data is parsed into data fields using a strictly enforced format scheme where similar data from CSES sources may be in differing formats. For example, the company name might be entered as ABC, Inc.; ABC Gum Company; or ABC Company. A translation lexicon would be used to parse the data into the desired format.

Understandably, people are nervous when data appears to change. However, there are three important points to bear in mind:

- The data in CSES remains untouched.
- The data is changed to more accurately reflect the truth (such as when the employer's company name is corrected) or to make the truth more apparent (such as when the occupation is standardized).

You control all transformations. As your organization develops the Central Metadata Repository (CMDR), you have the ability to control how each transformation will occur.

Step 7 shows the output of the NCP data dimension records. The most noticeable change is the addition of a unique record number.

Step 8 represents the CMDR, which enforces all business rules and ensures that all data elements are correctly formatted and drawn from the correct source and that all transformations are performed correctly. Most important, the CMDR ensures that a common language exists among all data marts and the analytical tools attached to them.

Step 9 is the data mart:

- It is a specialized database—specialized in the way data is stored compared with a typical transactional database. Still, you can implement a star schema for a data mart using any popular database package (Microsoft Access®, SQL Server®, Oracle®, Sybase®).
- The data mart is for storage. While many examples refer to using a data mart to develop queries and reports, it is actually the analytical tools, not the data mart, performing that work. Star schemas and a CMDR facilitate the interface between the data mart and the analytical tools. Without analytical tools, the data mart is just a collection of data with nothing to do.

Fact Table

The current support incentive measure requires two pieces of information:

- The Amount Collected for current support in IV-D cases (OCSE-157 Line 25)
- The Amount Owed for current support in IV-D cases (OCSE-157 Line 24)

Your SCS-DSS will use the field Current Collections to capture all collections—IV-D and those that are not IV-D.

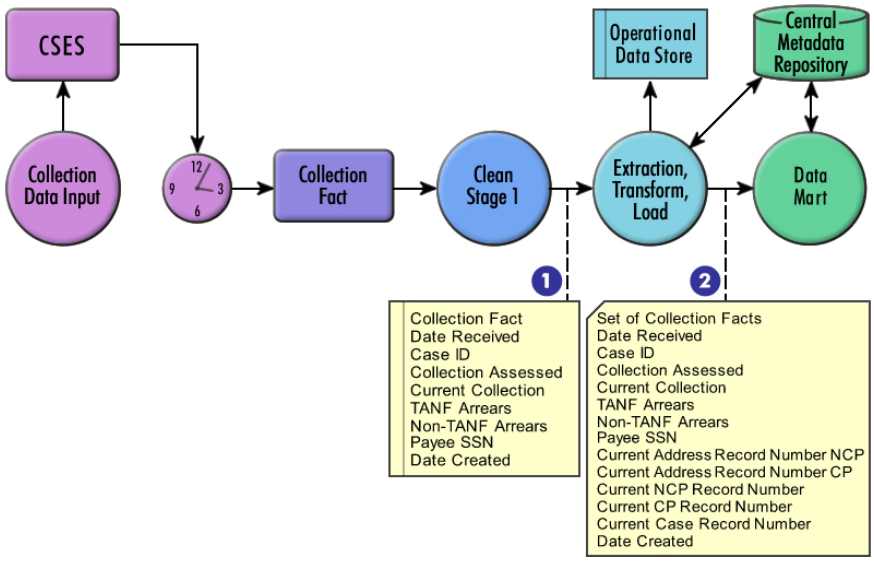

Figure 2-3. Fact Table Development Procedure

Information in the Case Data dimension indicates the type. Likewise, the field Collections Assessed refers to the collection amount due for the prescribed period of the support order (weekly, monthly, quarterly).

Let's presume that an NCP becomes unemployed and misses a payment in April 2001.

If the Collections Assessed field is used as a running tally, the report would look like this.

Date	Assessed Collections	Current Collections	TANF Arrears	Non-TANF Arrears
1/1/2001	$150.00	$150.00		
2/1/2001	$150.00	$150.00		
3/1/2001	$150.00	$150.00		
4/1/2001	$150.00	$0.00		
5/1/2001	$300.00	$300.00		
6/1/2001	$150.00	$150.00		
Totals	$1,050.00	$900.00		

Sum Current Collections	$900.00
Sum Current Collections Owed	$1,050.00
Collections Percentage	86%

However, since the Collections Assessed field is used only to indicate the current amount due, the data is actually reported like this.

Date	Assessed Collections	Current Collections	TANF Arrears	Non-TANF Arrears
1/1/2001	$150.00	$150.00		
2/1/2001	$150.00	$150.00		
3/1/2001	$150.00	$150.00		
4/1/2001	$150.00	$0.00		$150.00
5/1/2001	$150.00	$300.00		-$150.00
6/1/2001	$150.00	$150.00		
Totals	$900.00	$900.00		0

Sum Current Collections	$900.00
Sum Current Collections Owed	$900.00
Collections Percentage	100%

Developing a fact table is similar to developing dimension tables; however, there are some important differences.

First, the data in a fact table should be values that will undergo mathematical operations or statistical analysis. This is the prime reason that the field Collections Assessed sits in the fact table rather than a dimension table.

Second, the order in which fact table data is loaded into a data mart is crucial. Fact data should be loaded after all dimension data is updated. For example, say a Custodial Parent (CP) moved on June 1, 1999. If the fact table is loaded before the dimension tables are updated, the June 1999 payment may be applied to the wrong geographic dimension. When any data mart is developed, determining the timing and sequencing of the data loads will be a critical step.

The primary operation in the loading of the fact table is to properly associate it with the correct dimension records through the use of unique keys. In some instances, there may be a direct correlation between a fact field and a dimension table. The Date Received field is an example of this phenomenon. For other dimensions, you develop a unique key when the data dimensions are loaded. The SCS-DSS adds these unique values to the fact table during ETL.

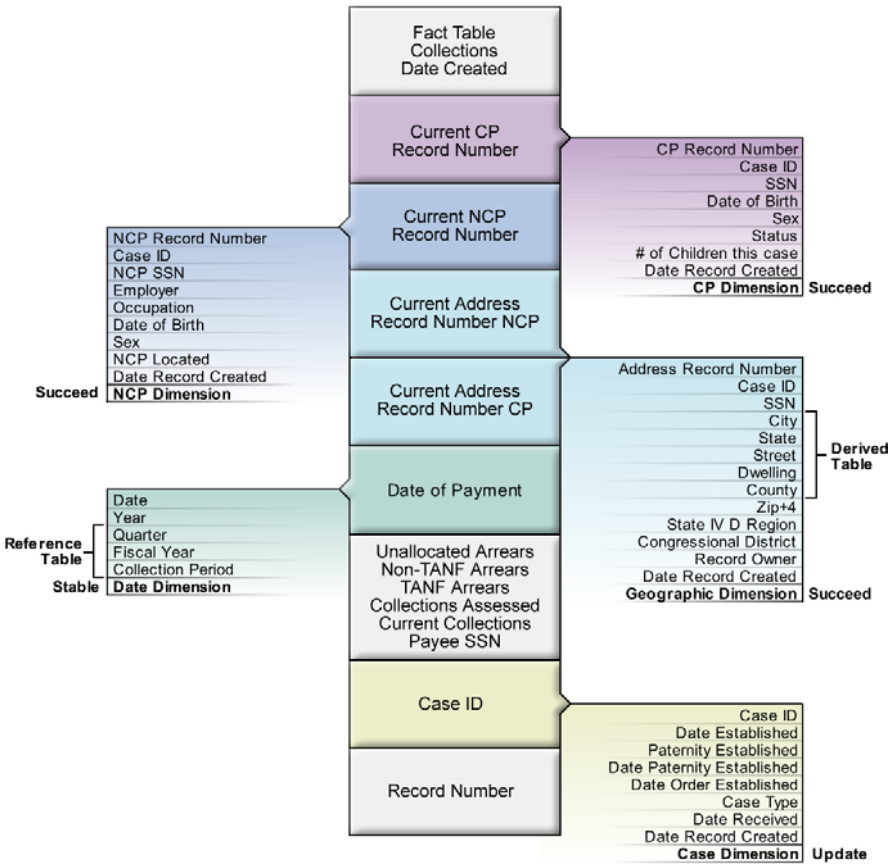

Figure 2-4. The Star Schema

- The Date dimension is referred to as stable because it will not be updated. For example, there will never be a need to change May 1, 2002, from a Tuesday to a Thursday. It is a reference table since most, if not all, star schemas within the data mart will use it.
- The Case Data dimension is an update table. It is unlikely that all of the data fields in this table will be filled in when a caseworker first opens a case. However, there should be one and only one value for each field over the life of that case. As the caseworker updates CSES, these fields will be updated in the SCS-DSS as well. This dimension links to the fact table using the Case_ID field.

One note: you may have at least one field that fails the “one and only one value” test. One common example is the Date Locate Initiated field. In real life, this action is likely to occur several times over the life of a single case. It was left in this version of the star to illustrate a point. During the design of your organization's SCS-DSS and its architecture, there will be many times when you think the design is complete. Either through testing, actual use, or design review, your team will discover a previously unconsidered relationship that will result in additional design work. This is normal and will continue throughout the life of the SCS-DSS.

- The NCP, CP, and Geographic dimensions are flagged with a succeed tag. This means that every time a data point changes, the SCS-DSS must receive a new dimension record to succeed the values in any previous record. Any fact records added after this time will be connected to the new dimension record; all previous fact records remain associated with the previous dimension record. This allows the SCS-DSS to review data over time.
- The Geographic dimension is a derived table. This means that information from CSES as well as other sources has been assembled within this record.

Concept of Operation

A data mart, no matter how well-designed, is just a hunk of data unless it interfaces with analytical tools. Consider the situation if a custodial parent (CP) moved from Nside of State to Sside of State on June 1, 1999. The case was transferred from Caseworker A63 in Office A to Caseworker B25 in Office B. Since 1997, the CP has received $100 per month in child support. Your regional management wants to know the collections for FY99 (Oct. 1, 1998, through Sept. 30, 1999), by quarter and by each of the region’s four offices. A caseworker enters this query into the SCS-DSS:

Select all Collection Records where Fiscal Year equals 1999. Sum Current Collections by quarter. Sum Current Collections by office.

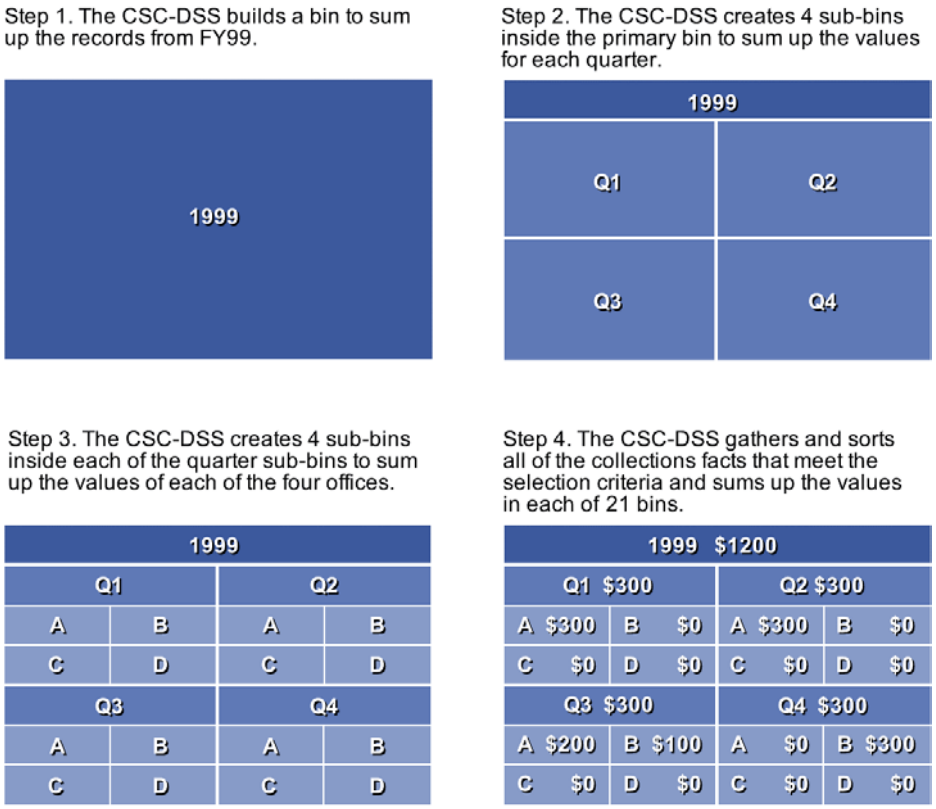

Figure 2-5. Example Binning

This is a simple example that indicates the power of an SCS-DSS. In a real-life implementation, it is possible to examine any fact or set of facts over any set or sets of dimensions to just about any breadth. The limiting factors are system speed, memory reserves, and the time/patience necessary to structure and execute the query. In a transactional database, such as CSES, the generally accepted performance standard is 2 seconds. This means that the system should return results within 2 seconds of a caseworker entering a case ID. With an analytical database, such as an SCS-DSS, the acceptable response time is 2 to 15 minutes, depending on the query's complexity.

Unfortunately, there are no CSE organizations in the real world with a single CP with an NCP who makes all child-support payments on time. To be effective, the SCS-DSS must be able to handle the vagaries of life for thousands of clients and still provide meaningful information. Toward that end, the next section will explore four key points:

- Currency (time)
- Granularity
- Sequencing
- Cardinality.

Currency

A data mart stores information over time so your organization can explore trends and patterns. If the collections data mart is going to provide geographic information about collections received from 1997 to 2001, it needs to know where everybody was in 1997, where they moved, and where they are now. Running a query without this information would provide you only the current address for each client.

The primary key to the Geographic, NCP, and CP dimensions in the fact table is the current Address, NCP, or CP record number. There may have been, of course, other records related to the same case. Consider the following example:

> J. J. Jerome, a custodial parent, enters the system on January 1, 1997. Monthly support payments of $100 began in 1997 and increased to $150 in 2000. All support payments have arrived on time. During

this period, Jerome has moved three times, each time within the state. Jerome came to the state IV-D office to request services, filed the necessary application, and paid the filing fee. Jerome has never received assistance. **Figure 2-6** shows two views of Jerome's support history.

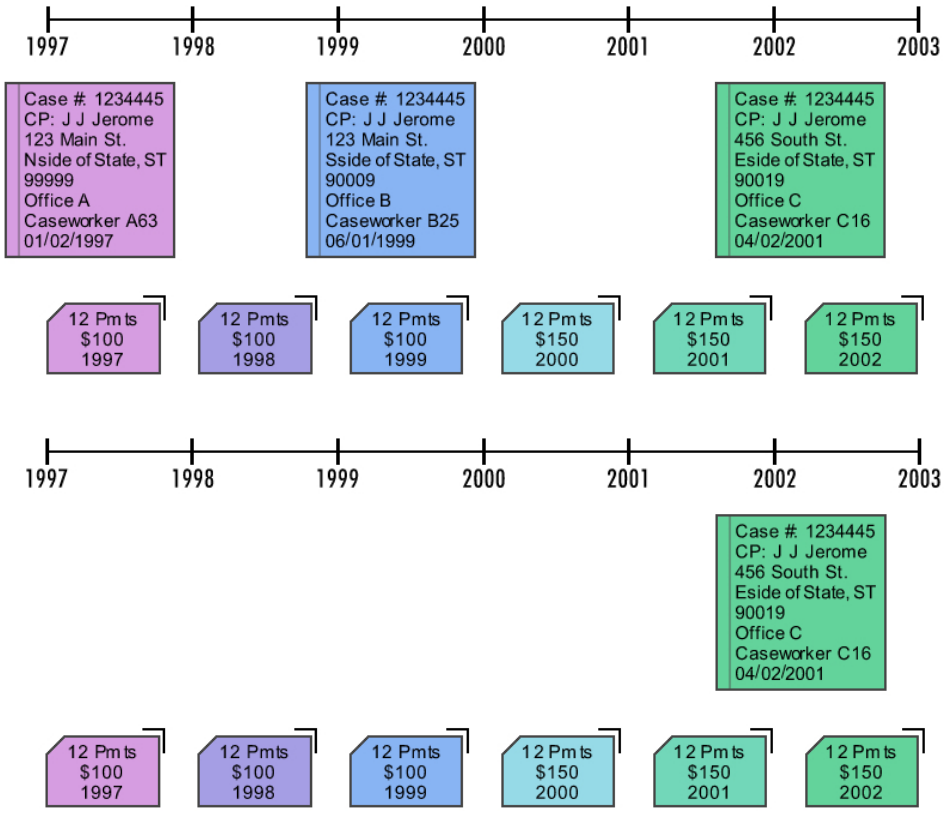

Figure 2-6. Two views of J. J. Jerome's Support History

The goal of the SCS-DSS is to maintain an accurate historical record of each case in the system so that it can provide the "current" truth for the period that is queried.

If only the current CSES information is transferred to the data mart, a request for all collections since 1997, sorted by city, would generate the following report (presuming Jerome is the only person in the system and that a complete set of payment records has been maintained).

Year	Locale	Amount	Year	Locale	Amount	Year	Locale	Amount
1997			1998			1999		
	Nside of State	$ -		Nside of State	$ -		Nside of State	$ -
	Sside Of State	$ -		Sside of State	$ -		Sside of State	$ -
	Eside of State	$1,200.00		Eside of State	$1,200.00		Eside of State	$1,200.00
	Wside of State	$ -		Wside of State	$ -		Wside of State	$ -
2000			2001			2002		
	Nside of State	$ -		Nside of State	$ -		Nside of State	$ -
	Sside Of State	$ -		Sside of State	$ -		Sside of State	$ -
	Eside of State	$1,800.00		Eside of State	$1,800.00		Eside of State	$1,800.00
	Wside of State	$ -		Wside of State	$ -		Wside of State	$ -

Figure 2-7. Jerome's support history based on current CSES data

When compared with the actual child support history, this is clearly incorrect. However, it was derived from the only information available. When designing a data mart, you need to make sure that you capture and incorporate a complete picture of the data. Some refer to this as a time slice; others call it a snapshot of the operational data. Depending on the SCS-DSS's ETL strategy, you will need to develop a procedure to move changes in key data fields to the SCS-DSS.

CSES may only maintain the current address; the SCS-DSS, on the other hand, will maintain all addresses. It does this by assigning a unique record number to each record placed into the Geographic dimension. It also enters a timestamp in the record to indicate when it was loaded into the data mart. As each collection fact is prepared for entry into the data mart, the system will associate the collection with the correct address. In computer pseudocode, it would look like this:

```
Find all geographic dimension records in which
Case_ID = Collection fact case ID AND
SSN = Collection fact payee SSN AND
Record Owner = NCP (This condition may be overkill)
Sort by Date Created Descending
(This ensures that the most current record is on the top of the list)
Find first Geographic dimension record (This directs you to the most current record)
Assign Record Number to Collection Fact Current Address Record Number NCP
```

The load process will actually perform a similar function for each dimension requiring the most current record. In the current star schema, this would include the Geographic, CP, and NCP dimensions. **Figure B-7** illustrates this process.

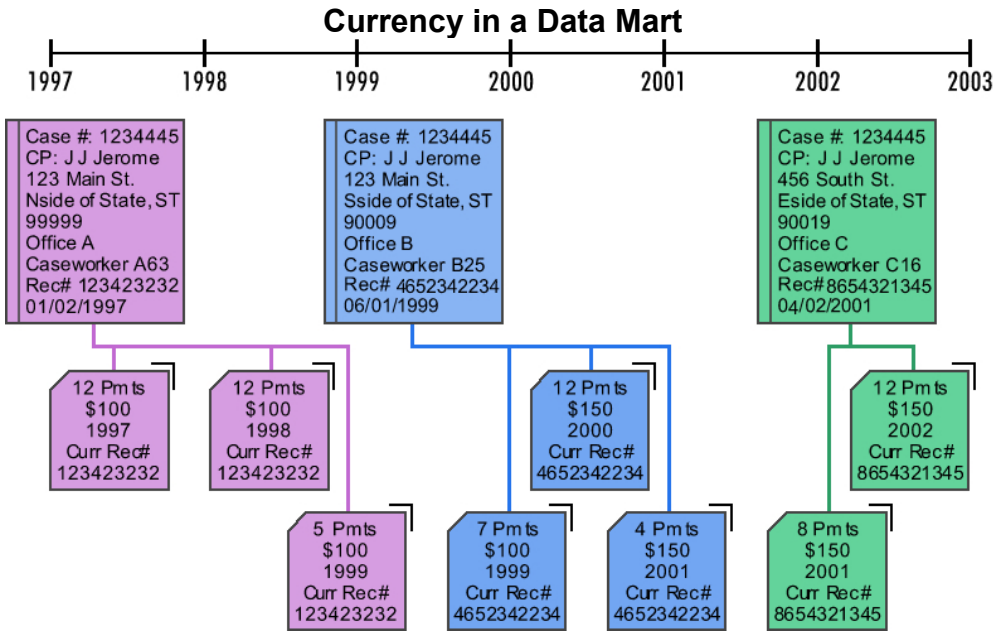

Figure 2-8. Jerome's Support History Allocated by Current Records

When the data is adjusted for currency, the same query now generates this table—the correct answer to the query:

Year	Locale	Amount	Year	Locale	Amount	Year	Locale	Amount
1997			1998			1999		
	Nside of State	$1,200.00		Nside of State	$1,200.00		Nside of State	$ 500.00
	Sside of State	$ -		Sside of State	$ -		Sside of State	$ -
	Eside of State	$ -		Eside of State	$ -		Eside of State	$ 700.00
	Wside of State	$ -		Wside of State	$ -		Wside of State	$ -
2000			2001			2002		
	Nside of State	$ -		Nside of State	$ -		Nside of State	$ -
	Sside of State	$1,800.00		Sside of State	$ 600.00		Sside of State	$ -
	Eside of State	$ -		Eside of State	$1,200.00		Eside of State	$1,800.00
	Wside of State	$ -		Wside of State	$ -		Wside of State	$ -

Granularity

Granularity represents the extent to which a system contains separate components (like granules). The more components in a system, the greater the granularity – and the more flexible it is. [1]

Another key factor is the level of detail. The more detail there is, the lower the level of granularity. The less detail there is, the higher the level of granularity.[2]

Confused?

Everyone seems to agree to agree that granularity is an important element in the design of a DSS, but no one seems able to agree on what it means. Rather than worry about trying to define it, let's examine the results.

Consider a retail store example: a chain store has just purchased a large quantity of blue parkas. Different functions operating within the organization will have different interests in the sale of these parkas:

- Personnel will be interested in who is selling them
- Marketing will be interested in where and when they are being sold
- Finally purchasing is interested in how many have been sold.

Depending on how data is loaded into the data mart, it may be possible answer all of these needs or only a subset of them.

If all the data from the point of sale is entered into the data mart, the blue parka sales table will look like this:

Item #	Description	Salesperson	Time of Sale	Date of Sale	Store	Qty Sold
356	Blue Parka	123	9:12	12/13/2000	16	1
356	Blue Parka	243	10:18	12/13/2000	34	2
356	Blue Parka	535	12:53	12/13/2000	106	1
356	Blue Parka	123	13:18	12/13/2000	16	3
356	Blue Parka	321	14:17	12/13/2000	43	2
356	Blue Parka	456	15:16	12/13/2000	45	1
356	Blue Parka	123	16:05	12/13/2000	16	3

With this data, it is possible to answer who, where, when (down to the minute), and how many.

If only daily sales totals are collected from each, the data will look like this:

Item #	Description			Date of Sale	Store	Qty Sold
356	Blue Parka			12/13/2000	16	7
356	Blue Parka			12/13/2000	34	2
356	Blue Parka			12/13/2000	106	1
356	Blue Parka			12/13/2000	16	2
356	Blue Parka			12/13/2000	43	1

It is possible to answer where, when (down to the day), and how many.

Finally, what happens if only product sales are captured?

Item #	Description			Date of Sale		Qty Sold
356	Blue Parka			12/13/2000		13

Now, it is only possible to answer the question when and how many.

The question that may be occurring to you at this point is why any organization would settle for less than the highest detail possible. The answer is speed, storage constraints, and business questions. In a real-world environment, a data mart may house millions of records occupying hundreds or thousands of gigabytes of space. When dealing with massive numbers, system performance and storage become very real constraints. Likewise, if

[1] http://www.webopedia.com/TERM/G/granularity.html

[2] W. H. Inmon, Building the Data Warehouse, 3rd Ed. (New York: Wiley Computer Publishing, 2002), p 43

the data mart is focused on a single set of functions, why incur the expense of buying a system significantly larger than what is required to do the job?

What does this mean in our world? Let's look at the system's records if data is transferred daily:

Daily Transfer						
			FACTS			
Client	Date	Change (dimension)	Current Collection	TANF Arrears	Non-TANF Arrears	Unallocated Arrears
J. J. Jerome	1/1/2000		$150			
J. J. Jerome	2/1/2000		$150			
J. J. Jerome	2/27/2000	Moved to Sside of State				
J. J. Jerome	3/1/2000		$150			
J. J. Jerome	4/1/2000		$150			
J. J. Jerome	4/23/2000	Status changed from never assisted to assistance				
J. J. Jerome	5/1/2000		$150			
J. J. Jerome	5/28/2000	Moved to Eside of State				
J. J. Jerome	6/1/2000		$150			

Queries posed to the SCS-DSS under these circumstances could generate a table of daily results. For example, a caseworker might ask the SCS-DSS to show collections from Jan. 1, 2000, to June 30, 2000, sorted by CP Status:

Month	Assistance	Never Assisted	Former Assistance	Collections
Jan-00		$150		
Feb-00		$150		
Mar-00		$150		
Apr-00		$150		
May-00	$150			
Jun-00	$150			

Here are the results if data is collected on a quarterly basis:

Quarterly Transfer						
			FACTS			
Client	Date	Change (dimension)	Current Collection	TANF Arrears	Non-TANF Arrears	Unallocated Arrears
J. J. Jerome	2/27/2000	Moved to Sside of State				
J. J. Jerome	3/31/2000		$600			
J. J. Jerome	4/23/2000	Status changed from never assisted to assistance				
J. J. Jerome	5/28/2000	Moved to Eside of State				
J. J. Jerome	6/1/2000		$600			

The same query would now generate a different report.

Month	Assistance	Never Assisted	Former Assistance	Collections
Jan-00				
Feb-00				
Mar-00		$600		
Apr-00				
May-00				
Jun-00	$600			

Sequencing

"When" is just as important as "what" when you are considering how to load a data mart. In transactional database processing (such as happens with CSES), the goal is to have the most current information. In an SCS-DSS, however, the goal is to have the most historically correct information.

Consider Jerome's move on June 1, 1999. How is the June 1 support payment attributed? Is it assigned to Office A since it could be considered the final payment for the third quarter of FY99? Or is it assigned to Office B since that staff had assumed responsibility for the case on June 1? For the sake of our example, we assumed the latter. During the data mart's design, determining the sequencing of data loading is second only to building the star schema design. Sequencing will be based on several factors, including current accounting and business practices, current closeout dates, and CSES availability.

During development and early operation of the SCS-DSS, you should select and monitor a small group of cases. These should be audited against a set of predefined reports and queries to ensure that the sequencing of data into the data mart is going as expected.

Cardinality

You are preparing a new dessert for dinner. You have all the ingredients, but time is a little short, so you press on. The recipe calls for a tablespoon of this, a tablespoon of that, and a teaspoon of salt. As you mix the dessert, you add a tablespoon of this, a tablespoon of that, and, most unfortunately for your family, a tablespoon of salt. You now have first-hand experience with the notion of violating cardinality.

For a DSS to function correctly, all data within a row of the fact table must be referring to the same level of detail. If the Assessed Collection equals a weekly or monthly payment, then any arrears accrual or payment must also be expressed in the same terms. This is important since the DSS expects to perform all the mathematical functions on the data contained in the fact table. If a running balance (total amount owed) is brought over for arrears each month, the value of uncollected will incorrectly soar.

When the Collections Data Mart is initially loaded, there will probably be a summary baseline set of facts and initial dimension values. The first fact record for J. J. Jerome's case may contain a summation of all transactions up to a specific point, such as before January 1, 2003. The next fact record for the case would reflect a monthly transaction.

Date	Assessed Collections	Current Collections	TANF Arrears	Non-TANF Arrears
12/31/2002	$9,000.00	$9,000.00	$0.00	$0.00
The next fact record would look like this				
01/01/2003	$150.00	$150.00	$0.00	$0.00